# 基于视觉注意模型的自适应视频关键帧提取

蒋 鹏 秦小麟
(南京航空航天大学计算机科学与工程系,南京 210016)

**摘 要** 关键帧提取是基于内容视频检索领域中一个重要的研究课题。提出了一种基于视觉注意模型的自适应视频关键帧提取方法。该方法分别提取视频中的运动和空间显著度,并用一种运动优先非线性混合模式将显著度合成为视觉注意度。在此基础上提出一种基于视觉注意度的局部和整体两级关键帧提取策略,先采用局部策略,选择镜头内注意度最大的帧作为关键帧候选;再根据视觉注意度的变化,为各个镜头自适应分配关键帧数目作为整体关键帧分配策略。实验证明,该方法提取的关键帧较为符合人类的视觉系统特性,而且该方法具有根据内容变化自适应提取关键帧等特点。

**关键词** 关键帧 视觉注意模型 自适应
**中图法分类号**: TP391.3 **文献标识码**: A **文章编号**: 1006-8961(2009)08-1650-06

# Adaptive Key-frames Extraction Based on Visual Attention Model

JIANG Peng, QIN Xiao-lin
(*Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016)

**Abstract** In this paper, we propose a novel video key-frame extraction method based on visual attention model. Firstly, the spatiotemporal saliency levels are generated and fused in a motion priority fashion to produce the overall attention degree. Then, a new adaptive key-frame extraction algorithm using attention and the variation of attention is put forward. For the shot level, the frames with higher attention value are selected as the candidates of the key-frames. For the clip level, the key-frame number is generated by the attention variation in a shot. Experimental results indicate the proposed method performs well in key-frame extraction with high efficiency.

**Keywords** key-frame, visual attention model, adaptive

## 1 引 言

随着数字电视以及多媒体技术的快速发展,产生了大量的视频文档。如何将海量视频数据进行有效的归类和管理是一个重要的研究课题。利用关键帧代表视频片段是一种常用的内容压缩方法,用户只需浏览少数能代表视频片段的关键帧就能快速了解视频片段内容。因此在该领域也出现了例如视频摘要(video summary)和镜头关键帧提取等应用。由于不同的视频在长度、内容等方面都有很大的区别,如何选择更具代表性和更能吸引观测者的视频帧作为关键帧是目前研究的一个重点问题。

在提取视频关键帧的早期工作中,一般采用全局的特征进行,例如采用颜色直方图或者运动直方图等方法对视频帧进行统计。一种简单的关键帧提取策略是利用镜头分割结果,采用镜头固定位置上的帧例如第 1 帧或中间帧为关键帧。但该类方法没有考虑到镜头内内容的变化,提取的关键帧代表性较差。基于全局特征的方法最主要的问题是将视频

**基金项目**: 国家自然科学基金项目(60673127)
**收稿日期**:2007-12-04;**改回日期**:2008-05-19
**第一作者简介**:蒋 鹏(1976~ ),男。南京航空航天大学计算机应用专业博士研究生。主要研究方向为多媒体数据库和计算机视觉。
E-mail:jpplus@163.com

帧作为一个整体进行考虑,并未考虑到视频中物体运动等高级语义特征。例如,在具有前景和背景的视频中,人们关注前景多于关注背景,背景的某些细微变化可能会导致颜色直方图或者运动向量的改变,从而提取的关键帧代表性较差。针对该问题,研究者们提出各种改进方法,例如文献[1]将视频中的物体进行分离,只考虑物体或感兴趣区域的变化而忽略背景等因素,这体现了人们视觉的注意性,相当于将人眼关注的部分进行了加权,因此获得了比较良好的效果。但该方法需要前景和背景足够明显且便于分割,而且对于没有明确物体的视频,该方法无法有效提取关键帧。近年来,出现了一种结合人类视觉系统(HVS)的视觉注意力模型,该模型根据人眼的视觉特性,将人眼感兴趣的注意区域通过视觉显著图表现出来。其中一种具有代表性的做法是 Itti 等人利用人眼对反差敏感特性,通过 Center-surround 算子将图像、视频中的显著区域提取出来形成一个合成的显著图(saliency map),显著图中由亮度值表示显著度,由强到弱依次表示显著性的强弱[2]。利用反差提取视觉注意区域的方法是一种自底向上(Bottom-up)的方法,该方法无需先验知识,也不需要根据具体视频内容设置或者调整视觉模型,因此受到了广泛的关注[3-4]。但是该方法是基于静态图像的,没有用到视频特有的运动特征,所以应用到视频中效果欠佳,而且该方法计算量较大,无法应用到实时性要求较高的应用中。因此研究者考虑将静态视觉注意区域和基于运动的动态视觉注意区域融合的方法。

文献[5]首次提出了利用视觉注意模型提取视频关键帧的方法,该方法分别生成静态和运动显著图,结合声音、人脸等特征分别产生一个 1 维注意度,并利用非线性混合方式产生注意度曲线,通过检测注意度曲线的区域极大值来检测关键帧。当注意度较高时,则认为该帧比其他帧更能吸引观测者的注意。该方法的优点在于利用了视觉注意模型,使提取的关键帧更符合人们的视觉感知认识,但对于注意度描述时,该方法仅依靠当前帧和周围几帧的局部信息,没有考虑到镜头甚至视频片段的全局信息,从而容易造成局部效应,有可能内容变化小的镜头反而比内容变化大的镜头提取更多的关键帧,虽然该方法通过设置最小提取长度来尽力避免该问题,但是该问题依然无法完全消除。

基于以上研究,提出了一种基于视觉注意模型的自适应视频关键帧提取方法。该方法分别提取视频中的运动和空间显著度,并用一种运动优先动态混合模式将显著度合成为视觉注意度。在此基础上,提出一种基于视觉注意度的局部和整体两级关键帧提取策略,先采用局部策略,选择镜头内注意度最大的帧作为关键帧候选。再根据视觉注意度的变化,为各个镜头自适应分配关键帧数目作为整体关键帧分配策略。其关键帧的数目是根据注意度的变化而自适应变化。这样有效地避免了某些长度长,但变化不太大的镜头产生过多的关键帧情况。

## 2　基于时空显著性的视觉注意度提取

视频可以看成是在时间轴方向变化的一系列图像,关键帧提取就是在大量的视频帧中提取最具有代表性的帧。视觉显著图的生成方法一般是将一个区域与周围区域进行比较,如果该区域与周围区域相似度越小,则该区域更能吸引注意力,也就是说该区域显著性较为明显。

### 2.1　运动显著区域提取

运动显著度是提取视频帧中各个块的运动向量,当运动向量较为一致时,其显著性较小,而运动向量不一致时,其显著度较大。例如:由摄像机平移造成的全局运动时,其运动一致性较好,而包含较多的局部运动时,其运动一致性较差。一般而言,观测者较为关注局部运动。因此,可以用运动的一致性来描述观察者对于运动的关注度。

首先将视频帧进行分块,块的大小为 8 × 8,设一帧图像中分成 $N$ 个块,则块描述符为

$$p_i(x_i, y_i, F_i, \mathrm{d}x_i, \mathrm{d}y_i), i \in N \tag{1}$$

其中,$x_i, y_i$ 为块中心位置,$F_i$ 为块的颜色,亮度等视觉特征。$\mathrm{d}x_i, \mathrm{d}y_i$ 为该块的运动向量。运动向量通过光流法计算获得。则第 $k$ 帧中第 $i$ 个块的运动方向为

$$\theta_i^k = \arctan\left(\frac{\mathrm{d}y_i}{\mathrm{d}x_i}\right) \tag{2}$$

游程为

$$\gamma_i^k = \sqrt{(\mathrm{d}x_i^2 + \mathrm{d}y_i^2)} \tag{3}$$

在运动方向和游程两个特征中,运动方向在运动一致性时是一个更具有描述性的特征,因此采用块的方向作为特征对于运动进行分析。考虑到邻近块运动的相关性以及为了减少对于参数的依赖,采用核密度估算(KDE)方法获得方向分布直方图,

KDE 定义如下：

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i) \tag{4}$$

采用高斯核函数进行核密度估计：

$$K(x, x^*) = \frac{1}{h\sqrt{2\pi}}\exp\left(-\frac{(x - x^*)^2}{2h^2}\right) \tag{5}$$

其中，$h$ 为核函数的带宽，在实验中，该值设置为 0.3。利用式(5)可将 KDE 统计看成一个直方图，最后用块 $i$ 与所有块的比值形式表示直方图中运动的一致性，运动注意度为

$$A_T(i) = \frac{\sum_{h=1}^{H} p(h) - p(h(i))}{\sum_{h=1}^{H} p(h)} \tag{6}$$

其中，$H$ 为直方图的槽数，而 $p(h)$ 为在 $h$ 处的概率，$h(i)$ 为块 $i$ 处对应的直方图槽数。通过式(6)，可以大致估算出视频中块 $i$ 与其他块运动分布一致性，首先把 $A_T$ 规格化到[0,1]，$A_T(i)$ 越大，表明块 $i$ 与其他块不一致性较大。为了消除静止或者细微运动块的影响，将游程加入到运动一致性计算：$A_T(i) = \theta(i) \cdot A_T(i)$，并计算所有块的运动一致性，构成 $A_T$。

## 2.2 静态显著区域提取

静态显著度提取是首先将容易引起观察者注意的区域提取出来，再根据显著性和位置等信息生成一个显著度。一般而言，当视觉显著区域位于屏幕中心，显著性较强时且面积较大时，则认为更能吸引人的注意力。

对于空间显著度，选择视觉较为敏感的亮度(I)、颜色(H)特征，进行静态显著区域检测，其中颜色特征由红绿对比和蓝黄对比结合而成。先采用数学形态学的开运算，即用 3×3 的结构元素先腐蚀再膨胀，消除噪声，再计算每个块的反差值。当一个块与周围的块之间反差值之和越大，则该块的反差越明显。块 $p_{i,j}$ 的反差值利用下式表示：

$$C_{i,j} = \sum_{q \in \Omega} d(p_{i,j}, q) \tag{7}$$

其中，$\Omega$ 为 $p_{i,j}$ 相邻的块，每一个相邻块由 $q$ 表示，而反差的距离度量为

$$d = \alpha \left| p_{i,j}^{I} - q^{I} \right| + \beta \left| p_{i,j}^{H} - q^{H} \right| \tag{8}$$

其中，$\alpha$，$\beta$ 为常量。在实验中，设置块的大小为 8×8，相邻块数目为 8。根据常识和一些研究表明[5]，处于屏幕中心的物体或者区域一般要比屏幕周围的区域更加重要，因此对于显著区域的空间位置关系进行加权形成静态注意度：

$$A_S = \frac{1}{N}\sum_{j=1}^{N} w_{x,y} \cdot C_{x,y} \tag{9}$$

其中，$w_{x,y}$ 为位置的权重值，定义参见文献[5]。

## 2.3 基于运动优先的视觉注意度生成

在提取了运动和静态显著区域后，将运动和静态显著区域合成一个视觉注意度，如下式：

$$A = w_S \cdot A_S + w_T \cdot A_T \tag{10}$$

其中，$w_S$，$w_T$ 分别为运动和空间显著度权重，研究文献[5]、文献[6]表明，采用固定权重值无法自适应根据视频内容变化调整运动和空间显著度比例。因此提出了一种运动优先的非线性混合模式将静态和运动显著度进行动态混合。当运动加强时，其运动权重迅速加大，静态图像权重迅速减小，但当运动强度达到一定强度时，其运动权重增加应该减缓。则运动优先原则定义为下式：

$$\begin{aligned} w_T &= \bar{A}_T \cdot \exp(1 - \bar{A}_T) \\ w_S &= 1 - w_T \end{aligned} \tag{11}$$

其中，$\bar{A}_T = \max(A_T) - \mathrm{mean}(A_T)$，即 $\bar{A}_T$ 为 $A_T$ 最大值减去 $A_T$ 平均值，能够自适应根据运动的反差变化调节 $\bar{A}_T$。$w_T$，$w_S$ 随 $\bar{A}_T$ 变化的曲线如图 1 所示。由图 1 可知，随着运动显著度的加强，运动权重值快速增加，同时静态权重值快速减少，当运动显著度达到一个值后，为了兼顾静态显著图，其 $w_T$ 增长速度减缓，而 $w_S$ 减少程度也减缓，这样的设计既考虑到运动优先，又同时兼顾静态部分。

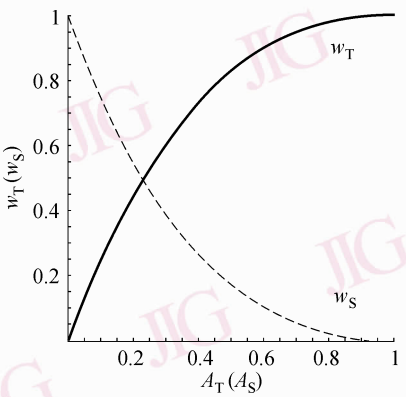



图 1 运动优先的动态权值变化图

Fig. 1 Motion priority scheme based static/motion weight curves

## 3　基于视觉注意度的关键帧提取

以上方法可以提取一个镜头内注意度最大的几个帧作为关键帧,但对于含有多个镜头的视频片段中每个镜头内提取关键帧的数目是不确定的。利用上述算法,一个变化很小的镜头,当把静态和动态注意度归一化后,可能会导致细微的变化都能表现出较大的曲线变化,从而导致提取更多的关键帧数目。由于视觉显著区域是人眼关注的区域,当视觉显著区域变化时,也就发生了视点转移,同时人的注意也发生了转移。基于上述观察,用一种根据视觉注意度变化方法来动态分配关键帧数目。如果一个镜头中,视觉注意度变化很小,也就是视点基本不转移,分配较少的关键帧,而当一个镜头视点转移较多,则分配较多的关键帧。对于一段视频,用户只需要设置一个关键帧的总数,该算法可以根据每个镜头的视点转移程度动态地为每个镜头分配关键帧数目。

首先设任意两帧之间视觉注意度的变化为

$$D_{i,j} = \left| \sum_{m=1}^{M} (A_m^i - A_m^j) \right| \tag{12}$$

其中,$M$ 为一帧中的分块数,则一个镜头的视觉注意度变化为

$$\bar{d} = \frac{\sum_{i=2}^{N} D_{i,i-1}}{N} \tag{13}$$

$N$ 为镜头视频帧数目。根据镜头变化给每个镜头分配的关键帧数目为

$$C = \max\left( T \frac{\bar{d}}{\sum \bar{d}}, 1 \right) \tag{14}$$

其中,$T$ 为给定的关键帧总数,每个镜头最少分配一帧作为关键帧。当镜头只有一个关键帧时候,则利用镜头内注意度最大的帧作为关键帧输出。为了避免关键帧在少数相邻的几帧中产生,导致产生的关键帧失去代表性问题,待提取关键帧到已经提取关键帧的距离 $D_{\text{key}}$ 需要满足下面不等式:

$$D_{\text{key}} > D_{\text{ave}} + \delta \cdot D_{\text{div}} \tag{15}$$

其中,$D_{\text{ave}}$ 为镜头内所有帧到已提取关键帧的距离的平均值,$D_{\text{div}}$ 为方差。$\delta$ 为一个常数,这里设为 1.5。在满足上述不等式的条件下选择注意度最大的帧作为关键帧。这样做的好处是一方面避免提取过于相似的关键帧,另一方面提取的关键帧具有较高的注意度。

## 4　实验结果及分析

为了验证本文算法的效果,选择了标准测试视频对基于视觉注意度模型的关键帧提取算法进行了验证。标准测试视频下载于 www. open-video. org。

第 1 个测试视频片段来自于 hcil2000_01. mpeg 中的第 2 个镜头,随着字幕的加入和人物的运动,镜头帧的静态、运动以及合并的注意度随之变化,其变化曲线如图 2 所示。

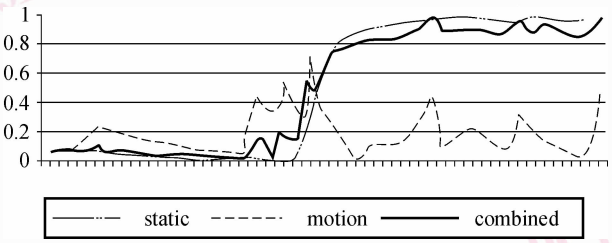



图 2　hcil2000_01. mpeg 视频片段空间,运动和合并视觉注意度变化曲线图

Fig. 2　Curves of static/motion/combined attention value from hcil2000_01. mpeg clip

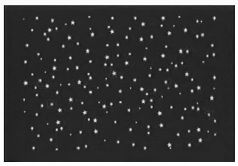
随着字幕的加入,空间显著度加大并保持稳定。而字幕的加入导致运动显著度也加大,由于人物的运动和背景噪声,导致运动显著度出现较大波动。当将空间和运动显著度合并以后,其视觉注意度大体由空间显著性决定,这是由于该镜头中运动较小造成的。图 3 为该镜头的关键帧选择比较图。


图 3　hcil2000_01. mpeg 视频中关键帧选取比较图

Fig. 3　Key-frame extraction comparison of hcil2000_01. mpeg clip

其中,(a),(b)为 2 个关键帧候选,图(c),(d)为图(a),(b)的视觉显著图。利用本文的基于视觉注意度模型,通过式(10)计算图(d)的视觉注意度为1 817 904,图(c)为1 176 944,本法选择注意度较高的帧作为关键帧,即选取图(b)作为关键帧。从图(a),(b)画面看,这样的选取符合人的视觉认知。

为了验证包含多个镜头的视频片段整体关键帧提取的效果,对 chi97_08_m1. mpeg 进行了测试,该视频总长 1 分 51 秒,共有镜头数 14 个。根据式(14)设置整个视频片段的关键帧总数,图 4 为该视频在关键帧总数为 21 时提取的关键帧,图中 S$x$F$y$ 代表第 $x$ 个镜头第 $y$ 帧。


S1F44

S2F149


S3F26


S4F99

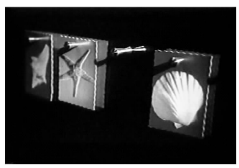

S5F2


S5F131


S6F1

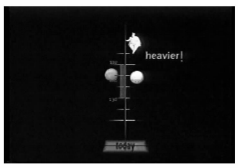

S6F58

S7F241


S8F20


S9F90

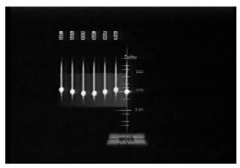

S9F278



S10F33


S11F19


S11F120


S12F142


S13F1


S13F13


S13F60


S14F1


S14F58

图 4 chi97_08_m1. mpeg 的关键帧提取结果

Fig. 4 Result of key frame extraction of chi97_08_m1. mpeg

从图 4 可以看出,当镜头相对静止时(镜头 1,2,3,4,7,8,10,12)该镜头内提取的关键帧数目少,提取的关键帧是该镜头内注意度最大的帧。而当镜头内容变化较大,如镜头 5,该镜头具有人物的运动和摄像机的缩进(zoom in),因此分配的关键帧数目较多。镜头 13 分配的关键帧数目最多,共有 3 帧关键帧,主要的原因是该镜头的运动较大,手从屏幕中迅速划过,因此根据运动优先原则,运

动权重加大,从而运动显著度也快速变化。同时该镜头伴有淡出(fade out)的镜头特效,其亮度渐渐变暗,造成静态显著图中块与周围块的反差减少,导致静态显著性变化较大。另外,从图 4 中可以看出同一个镜头提取的关键帧相似性较小,如 S6F1 和 S6F58,关键帧具有较好的代表性。这是由于本文方法在提取关键帧时同时考虑了视觉注意度最大和关键帧的距离两个因素,避免了关键帧在少数邻近且相似的几帧中产生。与文献[5]方法相比,本文方法可以根据整个视频片段内容变化,自动调整每个镜头输出关键帧的数目,有效地避免了镜头内容变化不大,但提取过多的关键帧问题。总体来看,关键帧提取结果符合人的主观判断,效果比较理想。

## 5　结　论

本文提出的基于视觉注意模型的关键帧提取算法,该算法分别提取空间和运动显著度,并采用运动优先原则将时空显著度进行动态混合形成视觉注意度。在一个镜头内提取注意度高的帧作为关键帧。为了解决关键帧数目的合理分配问题,用一种基于注意度变化的全局关键帧分配方案为镜头分配关键帧。实验结果证明,由于采用了符合视觉感知的注意模型,本文方法提取的关键帧较为符合人类的视觉系统特性。如何提取在语义层面上更有意义的关键帧是今后需要考虑的问题。

## 参考文献(References)


1　Liu Li-jie, Fan Guo-liang. Combined key-frame extraction and object-based video segmentation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2005,**15**(7):869-884.

2　Itti L, Koch C. Feature combination strategies for saliency-based visual attention systems[J]. Journal of Electronic Imaging, 2001, **10**(1):161-169.

3　Zhang Peng, Wang Run-sheng. Detecting salient regions based on location shift and extent trace[J]. Journal of Software,2004,**15**(6):891-898.[张鹏,王润生. 基于视点转移和视区追踪的图像显著区域检测[J]. 软件学报, 2004,**15**(6):891-898.]

4　Ansgar R. Koene,Li Zhao-ping. Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom-up saliency map in V1[J]. Journal of Vision,2007,**7**(7): 1-14.

5　Ma YF, Hua X S. A generic framework of user attention model and its application in video summarization [J]. IEEE Transactions on Multimedia,2005,**10**(7):907-919.

6　Yun Zhai, Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues[A]. In:Proceedings of 14th Annual ACM International Conference on Multimedia[C], Santa Barbara, CA, USA,2006:815-824.